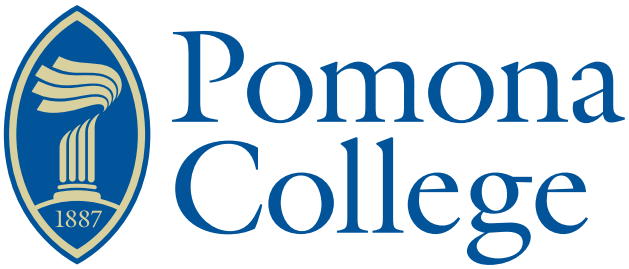


Senior Thesis in Mathematics

Trusting the Black Box: Confidence with Bag of Little Bootstraps

Author:
Christopher Garnatz

Advisor:
Dr. Jo Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 24, 2015

Abstract

In this paper we examine the Bag of Little Bootstraps, a method which yields an approximation for the sampling distribution of a statistic computationally. BLB is similar to bootstrapping but is optimized to better handle large datasets, and thus shows greater promise in a modern context. In this paper, we discuss the trade-off between interval accuracy and computational time, and we show that BLB with the proper hyperparameter values can return reliable intervals quickly.

Contents

Chapter 1

Introduction

Statistical inference aims to inform us about the value of a parameter for an unknown population, and to help us make sense of data that we may encounter in all areas of life. To make sense of data, we typically consider some quantity that is a function of the data of interest, a statistic. We then use the statistic as an estimator for the value of the parameter of interest.

We can think of a sample that we observe as a set of draws from an unknown population distribution. Given a sample, we seek to glean information about the larger population that includes the sample. Often we seek information about our population distribution in terms of the value of a parameter, a number which could describe the center or spread of the distribution, for example. But if the distribution is unknown, then the parameter value is unknown as well.

Fortunately, we can accurately guess at the true parameter values by using statistics, specifically the sampling distributions of statistics. For a given statistic, its sampling distribution describes all possible values and relative frequencies of the values that the statistic can take on, as the statistic is realized on all possible samples of a fixed size drawn from the population distribution. The sampling distribution is also unknown, but can be approximated.

If the parameter of interest is the population mean, then we can estimate the true value using the sample mean. The Central Limit Theorem tells us the sampling distribution is approximately normal, regardless of the underlying population distribution, provided that the sample size is sufficiently large. Thus, we can obtain an approximation for the sampling distribution without access to or knowledge of our population distribution.

But what if we are interested in the values of other parameters? We must choose a different statistic as an estimator for the parameter, but many statistics of practical importance lack the theory to describe the nature of their sampling distibutions independent of the underlying population distribution. This leaves us without an asymptotic approximation akin to the CLT.

Bootstrapping provides an alternative, computationally intensive route to approximate the sampling distribution of any statistic, for any population. Bootstrapping estimates the population distribution by using the single observed sample as an empirical distribution. To bootstrap, draw n observations from the original sample *with replacement* to obtain a single resample, then compute the value of the statistic for the resample. After a sufficient number of resamples have been generated, we aggregate the several computed values of the statistic to form a bootstrap distribution, an approximation for the sampling distribution.

Though the center of the bootstrap distribution will likely be biased in one direction relative to the true sampling distribution due to the randomness present in the approximate empirical population distribution, or the sample, the shape of the bootstrap distribution nicely matches the shape of the sampling distribution. Since we can correct for the bias present in our bootstrap distribution, we can use the bootstrap distribution to estimate the value of a parameter.

Bootstrapping provides an extremely powerful tool for statistical inference. Even if the theoretical shape for a sampling distribution is too difficult or impossible to derive theoretically, we can still approximate its shape computationally. With the knowledge that we can bridge the theoretical gap of sampling distributions, this thesis focuses on the computational challenges that bootstrapping encounters, especially for larger samples.

Chapter 2

Bootstrapping and the Bag of Little Bootstraps

2.1 Setting

We now introduce the setting and notation that will be used throughout this paper and adopt the notation in [4] to ease readability. We observe data $(X_1, .., X_n)$ where all observations are drawn independently and identically from an unknown population distribution P. Using our data, we can form the corresponding empirical distribution $\hat{P}_n$ for P as

$$\hat{P}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq t) \tag{2.1}$$

As the sample size n increases, the steps of the empirical distribution $\hat{P}_n$ increase in number and decrease in size, yielding a better approximation of the presumably smooth population distribution P. We now introduce our primary problem. We want to know the true value of a parameter θ for our population P, but we do not know P. Instead, we estimate θ with a statistic $\hat{\theta}_n$, the subscripted n denoting its reliance upon $\hat{P}_n$. Since our estimator is a function of random variables, our estimator is itself a random variable and thus has a distribution. We denote the theoretical sampling distribution of our estimator $\hat{\theta}_n$ as Q_n. Using our data and our estimator, we would like to obtain a subset of all possible values that the parameter could take where the true value of the parameter is likely to lie. This inference could come in the form of a quantile, bias, standard error, or confidence interval [4]. Let the

function $\xi(\cdot)$ denote the computation of the subset necessary for inference. For the duration of this paper, $\xi(\cdot)$ will compute the bounds of a confidence interval.

Our function ξ depends on the distribution of our estimator, and by extension, the distribution of our data. We denote this dependence by writing $\xi(Q_n, P)$. Our end goal is to compute $\xi(Q_n, P)$, but this quantity cannot be computed directly as P is unknown. Additionally, for most estimators $\hat{\theta}$, we do not know the sampling distribution Q_n. Thus we have two approximations to perform before we obtain a final estimate, one for our population P and one for our sampling distribution Q_n. The first approximation is straightforward, as we substitute the empirical distribution $\hat{P}_n$ given by 2.1 in for P. The Glivenko-Cantelli theorem justifies this substitution, as it proves that the empirical distribution $\hat{P}_n$ converges uniformly to P [5]. The second approximation requires more work to be obtained, but with bootstrapping, we obtain a data-driven approximation for the sampling distribution [2].

The data-driven approximation is constructed by taking r resamples of size n from our empirical distribution $\hat{P}_n$ and evaluating our statistic on each resample. We introduce the following notation to formalize this process: for $j \in \{1, 2, .., r\}$, let $X_j^* = \{X_{1j}^*, .., X_{nj}^*\}$ denote the jth resample with X_{ij} drawn i.i.d. from $\hat{P}_n$. Let $\hat{\theta}_{nj}$ be the value of $\hat{\theta}_n$ realized on the jth resample X_j^*. The empirical sampling distribution for $\hat{\theta}$ is then

$$\hat{Q}_n(t) = \frac{1}{r} \sum_{j=1}^{r} \mathbb{1}(\hat{\theta}_{nj} \leq t) \tag{2.2}$$

which estimates the unknown true sampling distribution Q_n. Combining 2.1 and 2.2, the bootstrap approximation for the desired quantity is now

$$\xi(Q_n, P) \approx \xi(\hat{Q}_n, \hat{P}_n) \tag{2.3}$$

where, as above, $\xi(\cdot)$ is a function which computes the endpoints of a confidence interval. Note $\xi(\hat{Q}_n, \hat{P}_n)$ represents the estimate of the CI based on the $\hat{Q}_n$, the estimate of the sampling distribution, and $\hat{P}_n$, the estimate of the population distribution.

2.2 Bag of Little Bootstraps (BLB)

2.2.1 Motivation

Bootstrapping provides us with a powerful tool for approximating quantities which have sampling distributions that are either too difficult or impossible to calculate analytically by creating an empirical sampling distribution for the statistic of interest. While powerful, the method's reliance on repeated resampling makes it computationally intensive and ill-suited for extremely large datasets. This section describes the Bag of Little Bootstraps (BLB), an alternative approximation for $\xi(Q_n, P)$ detailed in [4].

Traditional bootstrapping is ill-suited for large datasets. Bootstrapping requires resamples to be the same size as the original sample. Observations in resamples are drawn with replacement from the original sample, and though the number of repeated observations in each resample surely grows, the number of unique observations present in each resample increases as well. For samples of size n, if n is large, the number of unique observations present in each resample converges to $.632n$ [4]. In short, the amount of memory required to store each resample increases linearly with sample size.

A large number of unique observations in each resample increases the time needed to compute the value of a statistic on each resample. For quantile-based statistics, like the median, a time increase occurs because of the larger number of points necessary to sort. For statistics which make use of a weighted summation, like the mean, a time increase occurs because of the number of unique observations to be summed.

To alleviate the effects of longer statistic calculations, we could make use of a distributed computing hierarchy and implement parallelism to calculate the desired statistic of multiple resamples concurrently. Unfortunately, the large amount of memory required to store each resample introduces a new slowdown in a parallel framework. For a separate processor to calculate and evaluate a statistic for a single resample, the processor must send and successfully receive a request to access the necessary information in global memory, where the original sample is stored, or already have the resample completely stored in the processor's local memory.

A global memory made by the first processor will not accepted if a second processor is currently accessing any part of the memory which overlaps with the first processor's request. And as the amount of memory needed to store each resample is certainly larger than half the memory required to store the

original sample, as the resamples of a large sample of size n will contain around $.632n$ observations, a shared supply of global memory will like result in sequential, not concurrent, calculations.

To parallelize computation, the solution becomes to store each resample in the local memory of the processor computing the value of its statistic - instead of sharing one supply of memory, copy and distribute all necessary observations as needed. But sending memory between processors introduces a time cost that is significantly larger than increasing the number of computations. In short, navely parallelizing traditional bootstrapping may *slow* the algorithm for a huge sample.

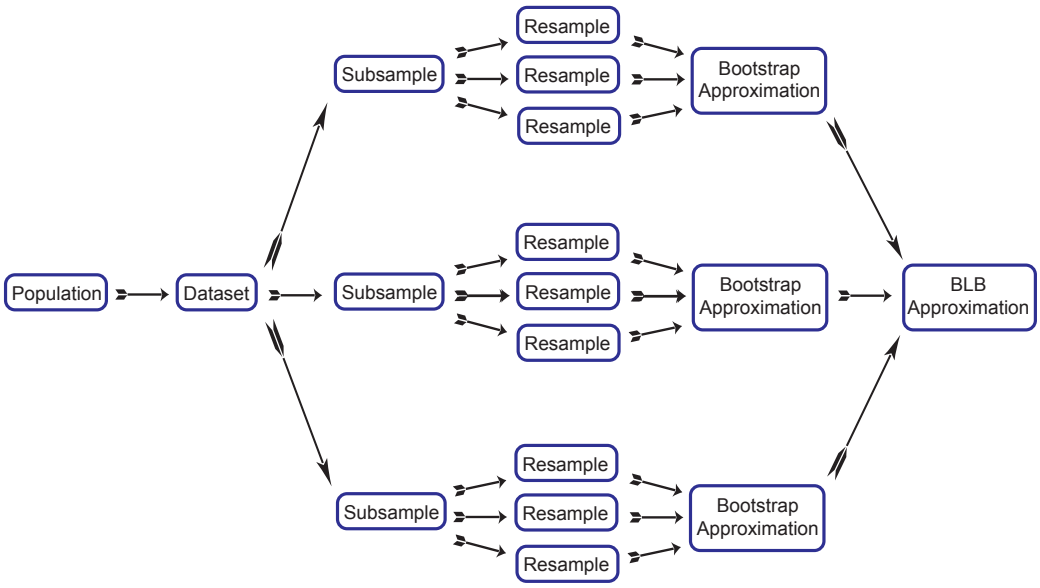


Figure 2.1: A flowchart showing the steps required in a BLB approximation

2.2.2 Description

BLB provides a powerful alternative to bootstrapping to approximate the sampling distribution of a statistic for a large sample. Unlike bootstrapping, BLB is designed for parallel computing, and enjoys a computational advan-

tage in handling large samples. BLB does this in two ways. First, it reduces the amount of information required to be sent between processors. Second, it increases the speed of computation of the statistic for each resample.

BLB approximates the sampling distribution of a statistic more efficiently than bootstrapping by calculating a number of inferior approximations for the sampling distribution, which are more readily obtained, and combining them. This process is depicted in Figure 2.1. Where bootstrapping draws its resamples from the original dataset, BLB instead draws its resamples from a number of subsamples, or subsets of the original dataset. Similarly, bootstrapping returns a single bootstrap approximation for the sampling distribution while BLB returns a bootstrap approximation for each subsample.

It should be noted that each bootstrap distribution computed for a subsample is an inferior estimation for the sampling distribution of our desired statistic than the standard bootstrap distribution. This is due to the fact that the subsample approximations contain fewer unique observations and thus less information than the bootstrap approximation whose resamples are drawn from the *entire* original sample. BLB overcomes this by averaging, and the results of these subsample approximations are then combined to create a more robust approximation. We will explore how different measures of central tendency affect the quality of the final BLB approximation.

Algorithm 1 Bag of Little Bootstraps (BLB)

Input: X, r, s, γ : $n = |X|$
Initialize: $b = n^{\gamma}$
for $k = 1, .., s$ **do**
  Subsample: $S_k \subset X$, $|S_k| = b$
  Population Distribution: $\hat{P}_{n,b}^{(k)}(t) = \frac{1}{b} \sum_{i \in \mathcal{I}_k} \mathbb{1}(X_i \leq t)$
  for $j = 1, .., r$ **do**
    Resample: $S_{k,j} \sim S_k$, $|S_{k,j}| = n$
    Statistic: $\hat{\theta}_{n,j}^{(k)*}$
  end for
  Sampling Distribution: $\hat{Q}_n^{(k)}(t) = \frac{1}{r} \sum_{j=1}^{r} \mathbb{1}(\hat{\theta}_{n,j}^{(k)*} \leq t)$
  Confidence Interval: $\xi(\hat{Q}_n^{(k)}, \hat{P}_{n,b}^{(k)})$
end for
Output: $\xi(\hat{Q}_n, \hat{P}) = \frac{1}{s} \sum_{j=1}^{s} \xi(\hat{Q}_n^{(k)}, \hat{P}_{n,b}^{(k)})$

2.2.3 Notation

We now formalize the BLB algorithm and will again adopt the notation in [4]. Let b be the subsample size, where $b = n^\gamma$. We consider values for $\gamma \in [0.5, 1)$. Let s denote the number of subsamples, or subsets of the original dataset. Finally, let r be the number of bootstrap resamples drawn from each subsample.

For $k = 1, .., s$, let $S_k \subset \{X_1, .., X_n\}$ be the kth subsample of size b of our data. Note that we sample from our data *without* replacement to create S_k. For each subset S_k, let $\mathcal{I}_k \subset \{1, .., n\}$ denote the corresponding set of indices for S_k, where $i \in \mathcal{I}_k$ if and only if $X_i \in S_k$. Thus $|I_k| = b$. We define the empirical distribution for the subset S_k as follows

$$\hat{P}_{n,b}^{(k)}(t) = \frac{1}{b} \sum_{i \in \mathcal{I}_k} \mathbb{1}(X_i \leq t) \tag{2.4}$$

We can now bootstrap from each of empirical distribution $\hat{P}_{n,b}^{(k)}(t)$. Let $\hat{Q}_n^{(k)}$ denote the bootstrap sampling distribution whose resamples are drawn from the empirical distribution $\hat{P}_{n,b}^{(k)}$. Though the s subsamples each only contain $b < n$ points, our r bootstrap resamples for each subsample will all be of size n. Having samples of size n allows the variance of the bootstrap distribution of the kth subsample, the variance of $\hat{Q}_n^{(k)}$, to be of the same order of magnitude as the variance of the sampling distribution for our statistic of interest, the variance of Q_n. There are alternative bootstrapping variants whose resamples are smaller than size n, but the approximate sampling distribution $\hat{Q}_n$ yielded by these methods must undergo analytical rescaling to match the variance of the variance of the true sampling distribution Q_n [1].

We now show how to construct $\hat{Q}_n^{(k)}$. We let k index over the subsamples and j index over the resamples. Let the multiset $X_j^{(k)*} = \{X_{1,j}^{(k)*}, .., X_{n,j}^{(k)*}\}$ denote the jth resample of the kth subsample. Note that $|X_j^{(k)*}| = n$ for all j, k. Let $\hat{\theta}_{n,j}^{(k)*}$ be the value of $\hat{\theta}_n$ realized for the jth resample of the kth subsample. We compile the realized values for the kth subsample into an empirical sampling distribution, defined as

$$\hat{Q}_n^{(k)}(t) = \frac{1}{r} \sum_{j=1}^{r} \mathbb{1}(\hat{\theta}_{n,j}^{(k)*} \leq t) \tag{2.5}$$

The end goal of BLB is to provide an approximation for $\xi(Q_n, P)$, a subset of the parameter space where the true value of our parameter is likely to

lie. Recall the bootstrap approximation for $\xi(Q_n, P)$ given by equation (2.3) substitutes the bootstrap approximation $\hat{Q}_n$ for Q_n. Similarly, the empirical sampling distribution for the kth subsample substitutes $\hat{Q}^{(k)}$ for Q_n. Since the resamples of the kth subsample were drawn from the empirical distribution of S_k given in (2.4), the approximation returned by the kth subsample is

$$\xi(Q_n, P) \approx \xi(\hat{Q}^{(k)}, \hat{P}_{n,b}^{(k)}) \tag{2.6}$$

as the variance of subsample's bootstrap distribution matches the variance of the true sampling distribution. BLB now has s approximations for $\xi(Q_n, P)$ and must combine them to yield a final approximation. The authors of [4] propose taking the mean of the k approximations. We also consider taking two other measures of central tendency, the median and interquartile mean, to average the approximations, but conclude from our experiments later in this paper that the mean yields the final approximation of the highest quality. This final approximation is given by

$$\xi(Q_n, P) \approx \frac{1}{s} \sum_{k=1}^{s} \xi(\hat{Q}_n^{(k)}, \hat{P}_{n,b}^{(k)}). \tag{2.7}$$

The final BLB approximation in (2.7) is better suited than the bootstrap approximation in (2.3) for parallelization. The subsamples of BLB attempt to minimize the amount of information that must be sent between processors, a significant computational cost. Once the information is distributed between multiple processors, the large number of resample statistics can be calculated simultaneously, not sequentially. By being better suited for parallelization, BLB is a better candidate to approximate the sampling distribution of a statistic for huge samples, as the increased number of computations can effectively be spread between multiple processors. Though the value of the statistic must be calculated for $s \cdot r$ resamples instead of r resamples, as a sampling distribution is generated for each subsample, the fewer number of unique observations present in each BLB resample decreases the amount of time needed for each calculation. Furthermore, the fewer number of unique observations present in each subsample decreases the amount of information sent between processors, significantly decreasing the runtime of BLB when the algorithm is implemented in parallel.

Chapter 3

Experiments

BLB provides an alternative data-driven method to bootstrapping for estimating the sampling distribution of a statistic. This alternative is especially useful as the sample size increases, as BLB is intelligently designed for its computational costs to scale nicely with massive datasets. However, this computational efficiency comes at the expense of a less accurate approximation of the sampling distribution.

The original paper on BLB [4] was published quite recently, in 2014. In the seminal work, the authors offer some advice on hyperparameter selection to obtain an accurate approximation. Setting all hyperparameters to sufficiently large values will yield an approximation that will surely be accurate, it will also involve extraneous calculations that will slow the algorithm's performance unnecessarily. We sought to understand the tradeoff between computational efficiency and estimation accuracy present in the algorithm's structure. The three hyperparameters of interest are

- γ : determines the size of the subsamples b, as $b = n^{\gamma}$
- s : determines the number of subsamples used
- r : determines the number of resamples drawn from each subsample via Monte Carlo simulation

We suspect that all three hyperparameters introduce a similar tradeoff to BLB's performance. If we increase the value of any of these three parameters, we increase the amount of information being used by the algorithm, and we should obtain a better approximation of our confidence interval. But

if the value of any hyperparameter is increased, we require either more memory or more processing power to accommodate the additional information, extending the running time of the algorithm. We wish to strike a balance between performance and efficiency and obtain an accurate approximation with minimal computation and memory use.

3.1 Bag of Little Bootstraps - Initial Tests

Our initial experiment sought to benchmark the approximation quality of BLB against a method with known theoretical guarantees. For this experiment, our sample was 10,000 i.i.d. observations from a standard normal distribution. Our end goal was to obtain a 95% confidence interval for the population mean. Our statistic for inference is the maximum likelihood estimator, the sample mean.

3.1.1 Experimental Setup

For $i = 1, .., 10000$, we observe $X_i \sim N(0, 1)$. Our sample is then $X = \{X_1, X_2, .., X_{10000}\}$. We are interested in the value of the population mean μ, but from our setup we know $\mu = 0$. We want to construct a 95% confidence interval for μ. We will empirically determine the confidence of our approximation by conducting 100 trials and calculating the proportion of times μ is included within the bounds of the interval.

To talk about this empirical confidence, we introduce a random variable Z. Let $Z_i = 1$ if the ith trial contains μ and let $Z_i = 0$ otherwise. We can then consider the proportion of times that Z will equal 1. Let $\pi = \mathbb{P}(Z = 1)$. We will estimate π with $\hat{p}$, where

$$\hat{p} = \frac{1}{100} \sum_{i=1}^{100} Z_i \tag{3.1}$$

because we conduct 100 trials. From the central limit theorem, we know that $\hat{p} \sim N(\pi, \sqrt{\pi(1-\pi)/n})$. Thus if we want a 95% confidence interval, $SE(\hat{p}) = \sqrt{\pi(1-\pi)/n} = 0.022$. Knowing this standard error, we deduce that for a confidence interval with a true level of 95%, or $\pi = 0.95$, we should expect to observe an empirical confidence level in the interval [90.6,99.4] around 95% of the time, given $n = 100$ trials. Thus for our BLB approximations, we

interpret empirical confidence levels of .91 to .99 as plausible measures for a true 95% confidence interval.

Since the parameter of interest is the population mean, we can rely on the central limit theorem to construct a 95% confidence interval with theoretical asymptotic guarantees. The benchmark that we use is a symmetric student-t confidence interval, or

$$[L, U] = [\bar{X} - t_{.975;n-1} \cdot \frac{s}{\sqrt{n}}, \bar{X} + t_{.975;n-1} \cdot \frac{s}{\sqrt{n}}] \tag{3.2}$$

where s is the sample standard deviation, $\bar{X}$ is the sample mean, and t is the .975 quantile of the t distribution with $n-1$ degrees of freedom. In addition to this benchmark, we also wish to compare the approximation quality of BLB and bootstrapping. We form an empirical distribution $\hat{P}_n$ as in (2.1) to estimate P. Next, we repeatedly take resamples of size n with replacement from $\hat{P}_n$, compute the value of our desired statistic on each resample, then form a bootstrap distribution of our desired statistic $\hat{Q}_n$ as in (2.2). From this, we can construct a 95% confidence interval, or evaluate $\xi(\hat{Q}_n, \hat{P}_n)$, in two different ways. The first method centers the interval about the mean of the empirical distribution and extends a specified number of standard deviations in each direction, forming a symmetric interval. The second method takes quantiles directly from the empirical distribution $\hat{Q}_n$. We consider both.

We can form different BLB confidence interval variants by altering two key steps in how BLB's approximation is formed. First, we can choose which type of bootstrap confidence interval is returned by each of the subsample empirical sampling distributions like we did with bootstrapping, obtaining $\xi(\hat{Q}_n^{(k)}, \hat{P}_{n,b}^{(k)})$, the bounds of a confidence interval returned by a single subsample's approximation, through either a symmetric bootstrap-t computation or through bootstrap quantiles. Second, we can choose how we aggregate the values $\xi(\hat{Q}_n^{(k)}, \hat{P}_{n,b}^{(k)})$ returned by each of the s subsamples. In (2.7), we took the mean of the upper and lower bounds to obtain a final pair of bounds. In this experiment, we also consider taking the median and the trimmed mean to obtain these final bounds.

In total, we empirically obtained an approximate confidence level for 9 different types of confidence intervals. They are

1. (std) Traditional inference on the entire sample: $(\bar{X} \pm t_{.95} \cdot s/\sqrt{n})$
2. (bst) Traditional bootstrapping: bootstrap-t CI

3. (bqt) Traditional bootstrapping: quantiles
4. (Btmn) BLB with bootstrap-t: mean of bounds
5. (Btmd) BLB with bootstrap-t: median of bounds
6. (Bttm) BLB with bootstrap-t: trimmed mean of bounds
7. (Bqmn) BLB with quantile bounds: mean of bounds
8. (Bqmd) BLB with quantile bounds: median of bounds
9. (Bqtm) BLB with quantile bounds: trimmed mean of bounds

For the BLB variants, we are specifically interested in the relationship between the hyperparameters and the true confidence level, as a correct confidence level near 95% indicates that BLB has likely returned a quality approximation. We first fix our subsample size by setting $\gamma = .5$. To determine how the number of resamples and the number of subsamples affects the quality of the approximation returned by BLB, we isolate these two effects by fixing $b = sqrtn$ and either r or s while the other varies. For the first experiment, we fix the number of subsamples and set $s = 100$ while incrementing the resample number r from 10 to 200 by steps of 10. For the second experiment, we fix $r = 100$ while incrementing s from 10 to 200 by steps of 10.

We expect these changes to have no effect on the student-t intervals, std, as these intervals not rely on r or s in any way. For the two types of intervals reliant upon traditional bootstrapping, bst & bqt, we expect a change in r to have an effect on their empirical confidence level while a change in s should have no effect at all. Because BLB relies upon the values of r and s for its final approximation, we expect to observe a change for the 6 BLB variants.

3.1.2 Results

The first experiment fixed the subsample size and number of subsamples and iterated over 20 different values of r for the number of resamples. As expected, the empirical confidence of the theoretically guaranteed student-t confidence interval fluctuates slightly around the true confidence level of 0.95, seen in the std column of Table 3.1. The two bootstrap intervals, bst and bqt, appear to generate high quality approximations, as their empirical

confidence also fluctuates around 0.95 for $r \geq 40$. As expected, the quantile-based approximation fails for the lowest values of r, as the .025 and .975 quantiles are much more prone to fluctuate in a distribution with only 10 or 20 observations (10,bqt), (20,bqt).

The six BLB variants fail to show the same quality as the bootstrap variants in their interval approximations in Table 3.1. Only two observations, (70,Btmn) & (60,Bqmn) exhibit an empirical confidence level of at least 0.9, and many observations are well below this threshold. We also see a general increase in confidence level for the three quantile-based BLB variants, though this increase never yields sufficiently close confidence measures ot 0.95. The improvement in BLB quantile performance happens for the same reason that the bootstrap approximation with quantiles improves as the number of resamples increases: with more observations in each BLB subsample's estimated sampling distribution $\hat{Q}_n^{(k)}$, we receive a more accurate approximation of the .025 and .975 quantiles of the true sampling distribution, Q_n.

The second experiment fixes the number of resamples and the subsample size and varies the number of subsamples, s. As expected, we again see the empirical confidence level of the theoretically guaranteed symmetric student-t confidence interval fluctuate about the true 0.95 confidence level. This is seen in the std column of Table 3.2.

We do not observe any trends in the confidence level given by the two bootstrap intervals, as expected. The only hyperparameter relevant to bootstrapping is the number of resamples r, which is fixed at $r = 100$. Since r appears to be sufficiently large to yield an approximation of decent quality, in Table 3.2, we see the confidence levels for bst and bqt fluctuate about .95 as well.

For the six BLB variants, subsample number appears to have a greater effect on the true confidence level of their approximations. For $s = 10$, the six BLB variants in Table 3.2 exhibit confidence levels of less than 50%. By contrast, when $s = 200$, the observed levels sit just below the true confidence level of 0.95, ranging from 0.87 to 0.94. From the trend of increasing confidence observed in Table 3.2, we conclude that number of subsamples, s, affects the quality of a BLB approximation more than the number of resamples, r.

One phenomenon present in Table 3.1 and Table 3.2 is the consistent underperformace relative to the targeted 95% confidence level, especially for lower values of the hyperparameters. This underperformance is due to the greater amount of variance present in a BLB approximation for smaller

r	std	bst	bqt	Btmn	Btmd	Bttm	Bqmn	Bqmd	Bqtm
10	0.98	0.93	0.83	0.81	0.81	0.84	0.72	0.64	0.72
20	0.94	0.9	0.84	0.84	0.72	0.8	0.76	0.77	0.7
30	0.94	0.93	0.88	0.83	0.75	0.75	0.77	0.73	0.75
40	0.96	0.98	0.91	0.81	0.71	0.76	0.76	0.78	0.75
50	0.97	0.94	0.93	0.79	0.78	0.77	0.78	0.74	0.74
60	0.98	0.96	0.97	0.81	0.77	0.83	0.9	0.77	0.71
70	0.96	0.95	0.94	0.91	0.77	0.78	0.75	0.79	0.72
80	0.97	0.97	0.95	0.84	0.8	0.84	0.78	0.76	0.82
90	0.95	0.94	0.94	0.78	0.75	0.88	0.86	0.79	0.85
100	0.94	0.93	0.93	0.82	0.79	0.78	0.83	0.77	0.76
110	0.93	0.89	0.89	0.73	0.74	0.76	0.8	0.72	0.73
120	0.94	0.95	0.91	0.78	0.83	0.82	0.85	0.76	0.82
130	0.94	0.93	0.89	0.8	0.71	0.75	0.82	0.72	0.78
140	0.95	0.93	0.91	0.77	0.76	0.76	0.8	0.73	0.79
150	0.96	0.95	0.94	0.76	0.75	0.84	0.69	0.74	0.81
160	0.94	0.95	0.97	0.86	0.77	0.74	0.83	0.82	0.78
170	0.97	0.94	0.95	0.8	0.7	0.84	0.74	0.7	0.78
180	0.97	0.98	0.95	0.83	0.76	0.84	0.77	0.77	0.78
190	0.98	0.98	0.97	0.86	0.81	0.82	0.83	0.74	0.85
200	0.94	0.94	0.94	0.83	0.78	0.79	0.81	0.75	0.84

Table 3.1: Approximately true Confidence Interval performance for bootstrap and BLB methods, benchmarked against student-t CI. $n = 10000$, $iters = 100$, $\gamma = .5$, $s = 100$, $r = [10 : 10 : 200]$

s	std	bst	bqt	Btmn	Btmd	Bttm	Bqmn	Bqmd	Bqtm
10	0.98	0.98	0.93	0.45	0.41	0.37	0.49	0.34	0.44
20	0.94	0.93	0.91	0.59	0.57	0.54	0.47	0.45	0.47
30	0.96	0.96	0.96	0.71	0.54	0.64	0.65	0.5	0.55
40	0.96	0.96	0.93	0.69	0.64	0.66	0.7	0.61	0.67
50	0.93	0.91	0.91	0.63	0.67	0.76	0.72	0.66	0.71
60	0.93	0.91	0.94	0.73	0.64	0.79	0.7	0.65	0.7
70	0.93	0.94	0.91	0.78	0.7	0.74	0.72	0.63	0.76
80	0.95	0.95	0.92	0.8	0.69	0.8	0.75	0.69	0.75
90	0.95	0.94	0.93	0.84	0.7	0.83	0.81	0.8	0.83
100	0.95	0.93	0.93	0.89	0.84	0.77	0.79	0.78	0.74
110	0.97	0.96	0.92	0.84	0.84	0.84	0.9	0.72	0.78
120	0.97	0.96	0.96	0.9	0.75	0.88	0.82	0.75	0.79
130	0.95	0.95	0.93	0.87	0.83	0.89	0.83	0.73	0.78
140	0.94	0.93	0.93	0.89	0.81	0.86	0.82	0.81	0.83
150	0.92	0.93	0.9	0.83	0.81	0.85	0.77	0.75	0.78
160	0.96	0.95	0.94	0.9	0.85	0.86	0.81	0.86	0.84
170	0.91	0.9	0.91	0.85	0.87	0.9	0.85	0.85	0.78
180	0.97	0.97	0.96	0.92	0.86	0.87	0.89	0.89	0.88
190	0.92	0.92	0.93	0.9	0.81	0.89	0.85	0.81	0.85
200	0.98	0.97	0.97	0.91	0.92	0.94	0.87	0.87	0.91

Table 3.2: Approximately true Confidence Interval performance for bootstrap and BLB methods, benchmarked against student-t CI. $n = 10000$, $iters = 100$, $\gamma = .5$, $r = 100$, $s = [10 : 10 : 200]$

hyperparameters. With fewer resamples, the bounds of the approximate confidence interval returned by each subsample exhibit greater variance. If these bounds move away from the mean and the interval is too large, this effect is not seen immediately. By contrast, if one bound strays too far in the other direction, a more likely scenario with increased variance, the intervals fails to capture the population mean more than it should. The phenomenon of these failures, called a miss percentage, are discussed at length in [3]. This effect is mitigated by increasing the number of subsamples, and thus increasing the number of confidence interval approximations to average over. In short, too small values for the hyperparameters of BLB will yield an α-level confidence interval whose true confidence is less than α.

3.2 Approximation Quality

The primary benefit of the BLB algorithm is its comparatively fast runtime when handling massive datasets. Unfortunately, increasing the number of subsamples or resamples increases the amount of memory and number of computations required to perform a BLB approximation, and we wish to optimize performance by minimizing extraneous calculations and memory usage. We are left with a tradeoff between approximation quality and computational performance.

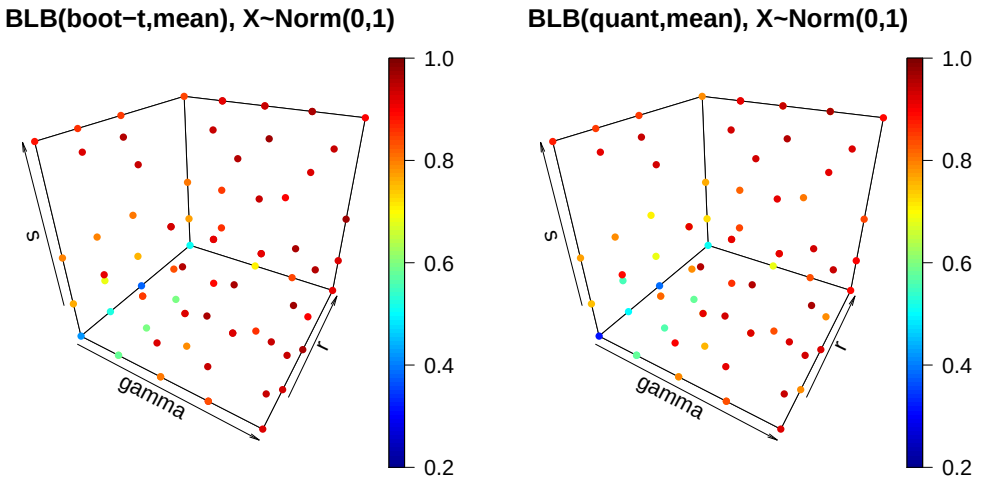
Our initial tests enlightened us to the fact the hyperparameter values do affect the approximation quality provided by BLB. We saw that the number of resamples affected the quality of the approximation somewhat in Table 3.1, but this effect was less than the number of subsamples observed in Table 3.2. After observing this effect on approximation quality, we conducted a more thorough grid search of the hyperparameters of BLB, calculating the empirical confidence level of BLB for different values of r, s, and γ. From this grid search, we hope to paint a clearer picture of the relationship between hyperparameter value and approximation quality.

We conducted two additional experiments for this grid search, creating 95% confidence intervals for the population mean. The first experiment generated data from a standard normal distribution; the second experiment took draws from an exponential distribution. For both experiments, each sample consisted of 10,000 observations, and 100 samples were observed.

After observing the performance of the six BLB variants in Tables 3.1 and 3.2, we concluded that taking the mean of the subsample approximations, given in 2.4, did not produce an inferior quality approximation than taking the mean of the subsample approximations. Thus our BLB approximation for these experiments take the form the authors of [4] suggest, given by equation (2.7). For each experiment, we still produce two type of BLB approximations. The first has the subsamples produce symmetric bootstrap-t confidence intervals; the second has the subsamples construct quantile-based intervals. In Tables 3.1 and 3.2, these two variants are Btmn and Bqmn, respectively.

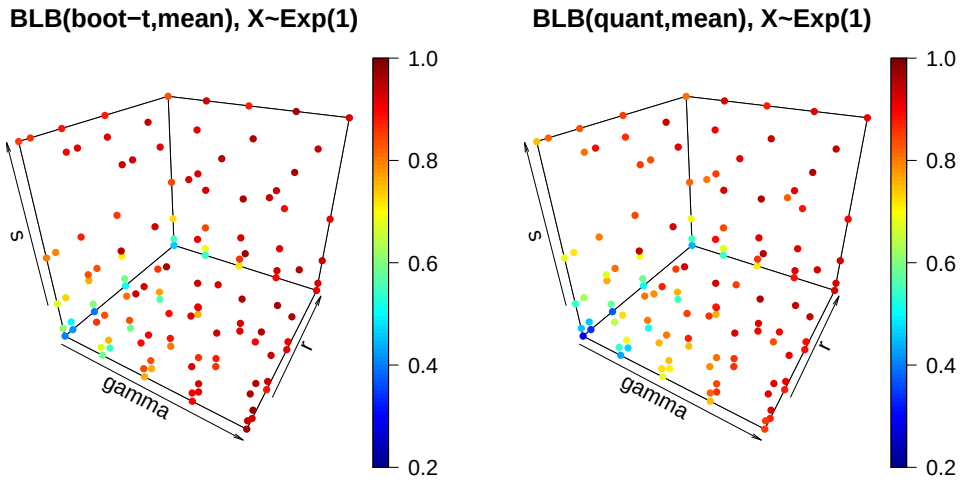
For our first experiment with normal data, we chose the following values to grid search over:

- $\gamma \in \{.5, .6, .7, .8, .9\}$
- $r \in \{10, 50, 100, 200\}$



(a) (b)

Figure 3.1: Standard Normal Data, BLB true recovery percentage of 95% confidence intervals constructed by taking the mean of the bounds of the subsamples' intervals which were formed by symmetric bootstrap-t distribution (a) or by bootstrap quantiles (b)



(a) (b)

Figure 3.2: Exponential Data, BLB true recovery percentage of 95% confidence intervals constructed by taking the mean of the bounds of the subsamples' intervals which were formed by symmetric bootstrap-t distribution (a) or by bootstrap quantiles (b)

- $s \in \{10, 50, 100, 200\}$

In Figure 3.1, we see that increased values of γ and s improve empirical confidence more than an increase to r. This is seen by beginning at the origin in (a) and (b) in the lower left-hand corner, then looking at the color of the dots along the three axes emanating from the origin.

For our second experiment with exponential data, we chose the following values to search over:

- $\gamma \in \{.5, .6, .7, .8, .9\}$
- $r \in \{10, 20, 50, 100, 200\}$
- $s \in \{10, 20, 50, 100, 200\}$

In Figure 3.2, we see the hyperparameters have the same effect on BLB approximation quality as they did in the first experiment with normal data. The lowest values for the three hyperparameters yields the worst approximation, and the quality of the BLB approximation improves more when the number and size of subsamples are increased than when the number of resamples is increased.

Taken together, Figures 3.1 and 3.2 show that the subsample size and subsample number, hyperparameters γ and s, exhibit a larger marginal improvement in the quality of the final BLB approximation than resample number, r.

3.2.1 Increased Detail

Figures 3.1(a) and 3.2(a) provide a enlightening visualization for how the three hyperparameters affect the approximation quality of a BLB confidence interval approximation using symmetric bootstrap-t intervals. We reran the two experiments - computing the empirical confidence level of a BLB approximation - that produced figures 3.1 and 3.2 and iterated over more values of r, s and γ.

For the first rerun, we generated data from a standard normal distribution and iterated over the following hyperparameter values:

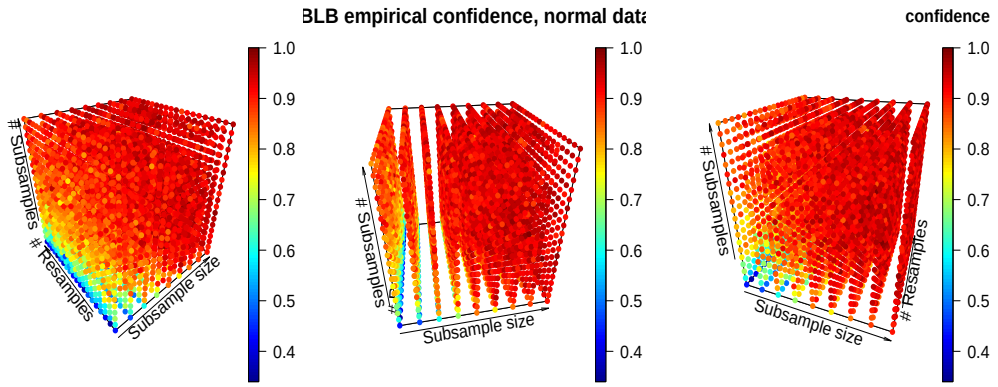
- $\gamma \in \{.5, .55, .., .9\}$
- $r \in \{10, 20, .., 200\}$

- $s \in \{10, 20, .., 200\}$

The empirical confidence levels from this first experiment are plotted in figure 3.3. For the second rerun, we generated data from a standard exponential distribution and iterated over the following hyperparameter values:

- $\gamma \in \{.5, .55, .., .7\}$
- $r \in \{10, 20, .., 100\}$
- $s \in \{10, 20, .., 100\}$

and plotted the empirical confidence levels in figure 3.4.

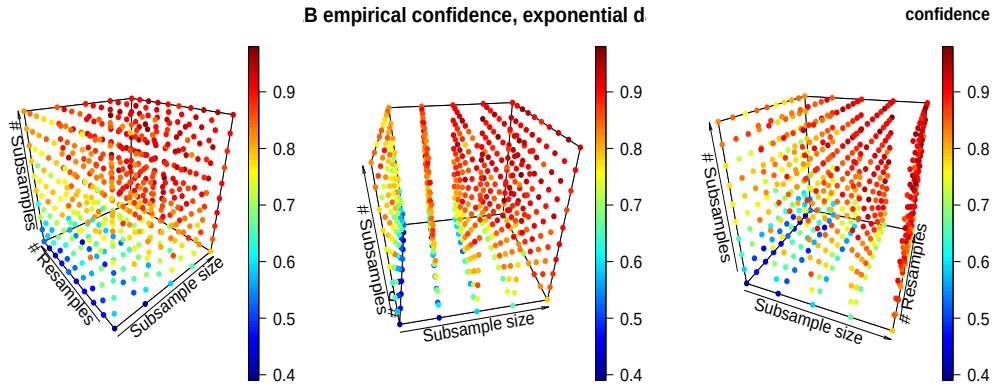


(a) (b) (c)

Figure 3.3: Three views of empirical confidence levels for approximate 95% confidence intervals for the population mean constructed using BLB on standard normal data. Empirical confidence levels for different values of r, s, and γ are shown. The bounds of each BLB subsample's interval were computed using the symmetric bootstrap-t method.

3.3 Approximation Efficiency

Our next step was to examine the effects that these three hyperparameters had on the overall runtime of the BLB algorithm. We've seen that s and



(a) (b) (c)

Figure 3.4: Three views of empirical confidence levels for approximate 95% confidence intervals for the population mean constructed using BLB on standard exponential data. Empirical confidence levels for different values of r, s, and γ are shown. The bounds of each BLB subsample's interval were computed using the symmetric bootstrap-t method.

γ have a larger effect on the quality of BLB's approximation, so we turn to examine whether these larger benefits to quality come at at a larger cost to runtime.

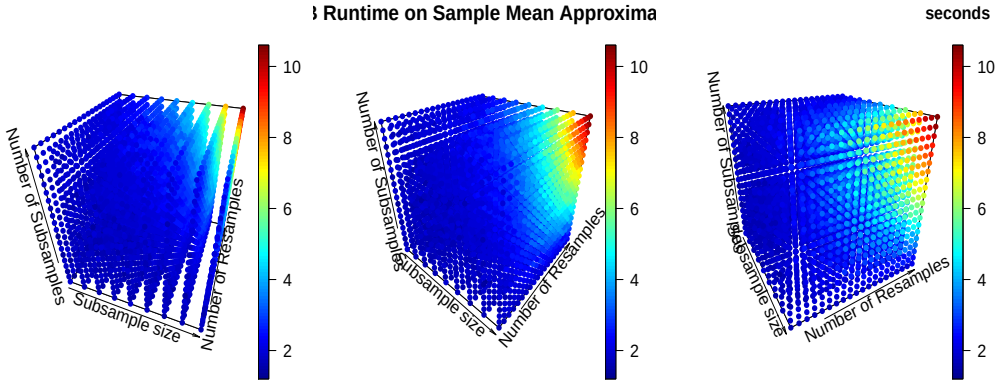
To better understand the quality/time tradeoff present in each hyperparameter, we conducted an experiment. For each trial of each set of parameters, we drew 10,000 observations from a standard normal distribution, computed a BLB approximation for a 95% confidence interval for the population mean, and logged the time that elapsed during the approximation. We chose the following values to search over:

- $\gamma \in \{.5, .55, .6, .., .9\}$
- $r \in \{10, 20, 30, .., 200\}$
- $s \in \{10, 20, 30, .., 200\}$

For each triplet of hyper parameters, we conducted 100 trials, and computed the mean of the 100 runtimes. The analysis was performed on a computer

with two eightcore AMD Opteron 6276 processors running at 1.4 GHz. The maximum number of allotted clusters was capped at 10.

We plotted the mean runtimes and show the plot viewed from three different angles in Figure 3.5. We immediately see that for most triplets of the hyperparameters, the runtime of BLB was quite fast. We see a jump in runtime in the upper-right hand corner of the three plots, especially for the largest subsample size where $\gamma = .9$. Figure 3.5(c) shows that increases in the number of resamples and subsamples also increase runtime, though less so than an increase in subsample size. We should recall that the subsample size is defined as $b = n^{\gamma}$, an important feature which ensures that the subsample size does not increase linearly with the sample size. This relationship also makes subsample size not increase linearly in Figure 3.5 as it does with γ as resample and subsample number do with r and s.



(a) (b) (c)

Figure 3.5: Three views of runtime for different values of r, s, and γ. Each trial saw BLB construct an approximate 95% confidence interval for the population mean given 10,000 observations from a standard normal distribution. Each point shows the average runtime of 100 trials for each triplet of r, s, and γ.

Chapter 4

Conclusion

Given a massive dataset, the Bag of Little Bootstraps (BLB) algorithm proposed by the authors of [4] provides an alternative to bootstrapping for approximating the sampling distribution of a statistic computationally. Crucial to the algorithm, BLB requires the inputs of three hyperparameters - γ, s, and r - to determine the size of subsamples, number of subsamples, and number of resamples, respectively. Inherently present in each of the three hyperparameters is a tradeoff between approximation quality and computational efficiency. If the triplet of parameters are too small, an unreliable approximation is returned. If the triplet of parameters is too large, extraneous calculations are made and BLB's primary advantage over bootstrapping for large samples, speed, is nullified.

In examining the quality of the BLB's approximation, empirically measuring the confidence level of a proposed 95% confidence interval, we discovered that subsample size and number affected BLB's approximation quality more than resample number. This effect is best seen in Figures 3.3 & 3.4. In both plots, we see the observed confidence is well below 95% percent for the minimum triplet of hyperparameters, each in the lower left-hand corner of the plot. If we fix r and s and move along γ's axis, we eventually obtain an interval whose empirical confidence level is close to 95%. A similar phenomenon is seen for s. But if we fix γ and s and move along r's axis, we do not obtain an approximate 95% confidence interval of any sort of quality.

In examining the runtime of BLB, we refer to Figure 3.5. The largest swath of red, indicating slower performance, mostly exists for the largest value of γ. From our runtime experiment, we concluded that a larger subsample size slows performance more than a larger number of subsamples or

resamples.

Synthesizing the results of our experiments to test BLB's quality and efficiency, we propose that the hyperparameter of immediate interest for those who wish to use BLB is s, the number of subsamples. If one is worried about the quality of an approximation for $\xi(Q_n, P)$ returned by BLB, one should first increase s, as s affects approximation quality of BLB more than increasing r while slowing the computation of the approximation less than increasing γ. We believe this point is of utmost interest, as anyone using BLB in practice would not build confidence intervals for the population mean or any other parameter of the population with theoretical guarantees. We would not have an interval with theoretical guarantees to compare our width of our approximate interval to. Furthermore, we cannot perform the same procedure that we did in out experiments to empirically estimate the level of our confidence interval, as we do not know the value of the parameter or have access to additional draws from our population. Thus to ensure quality, we should err on the side of extraneous calculations that may slow performance but deliver a sufficient approximation. An increased number of subsamples optimizes the tradeoff between quality and efficiency.

In our experiments, one factor that we did not take into account was sample size. It is not yet clear whether the affects the hyperparameters have on the quality and efficiency of the approximation scale in a similar manner as the sample size increases. Further research on the Bag of Little Bootstraps could address this topic as a starting point.

Bibliography

[1] P.J. Bickel, F. Gtze, and W.R. van Zwet. Resampling fewer than n observations: Gains, losses, and remedies for losses. In Sara van de Geer and Marten Wegkamp, editors, *Selected Works of Willem van Zwet*, Selected Works in Probability and Statistics, pages 267–297. Springer New York, 2012.

[2] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1994.

[3] Tim Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. arXiv preprint, November 2014.

[4] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. A scalable bootstrap for massive data. *J. R. Stat. Soc. B*, 76(4):795–816, 2014.

[5] Howard G. Tucker. A generalization of the glivenko-cantelli theorem. *The Annals of Mathematical Statistics*, 30(3):pp. 828–830, 1959.